Statistical Post-Processing of Ensemble Precipitation Forecasts by Fitting Censored, Shifted Gamma Distributions

Michael Scheuerer*

University of Colorado, Cooperative Institute for Research in Environmental Sciences at the NOAA Earth System Research Laboratory

Thomas M. Hamill

National Oceanic and Atmospheric Administration (NOAA), Earth System Research Laboratory, Boulder, Colorado

Corresponding author address: Dr. Michael Scheuerer, NOAA/ESRL, Physical Sciences Division, 325 Broadway, R/PSD1, Boulder, CO 80305-3337.

E-mail: michael.scheuerer@noaa.gov

ABSTRACT

We present a parametric statistical post-processing method which transforms raw (and frequently biased) ensemble forecasts from the Global Ensemble Forecast System (GEFS) into reliable predictive probability distributions for precipitation accumulations. Exploratory analysis based on 12 years of reforecast data and 1/8-degree climatology-calibrated precipitation analyses shows that censored, shifted gamma distributions can well approximate the conditional distribution of observed precipitation accumulations given the ensemble forecasts. A nonhomogeneous regression model is set up to link the parameters of this distribution to ensemble statistics which summarize the mean and spread of predicted precipitation amounts within a certain neighborhood of the location of interest, and in addition the predicted mean of precipitable water. Regression parameters are fitted to training data through minimization of the continuous ranked probability score. The proposed method is demonstrated with precipitation reforecasts over the conterminous United States using common metrics such as Brier skill scores and reliability diagrams.

1. Introduction

Ensemble predictions are now routinely generated at operational weather prediction centers worldwide (Molteni et al. 1996; Toth and Kalnay 1993, 1997; Houtekamer and Derome 1995; Charron et al. 2010). Despite many improvements to them over the last $\sim$2 decades, precipitation forecasts from the ensembles are still typically unreliable, be it from insufficient model resolution, less-than-optimal initial conditions, sub-optimal treatment of model uncertainty, and/or sampling error. For this reason, statistical post-processing of the output of an ensemble prediction system is commonly an integral part of the forecast process, since it can improve the reliability and skill of probabilistic guidance (e.g. Wilks and Hamill 2007; Hamill et al. 2008, and references therein). By comparing past forecasts with their verifying observations, systematic biases and inadequate representation of forecast uncertainty can be identified, and the current forecast can be adjusted such as to minimize these systematic errors. When the forecasts are provided on a grid that is too coarse to resolve small-scale effects that affect the weather variable under consideration, many post-processing methods also implicitly perform a statistical downscaling.

The statistical post-processing of precipitation accumulations is far more challenging than the post-processing of weather variables like surface temperature or wind speed for several reasons:

1. Their mixed discrete/continuous nature (positive probability of being exactly zero, continuous value range for positive precipitation amounts) makes it difficult to find an adequate parametric distribution model.

2. Forecast uncertainty typically increases with the magnitude of expected precipitation amounts; this must be taken into account when setting up a model for the conditional distribution of observed precipitation amounts given the ensemble forecasts.

3. High precipitation amounts occur very infrequently; a customized treatment of these cases may therefore require a vast amount of training data.

The advantages and disadvantages of the different post-processing approaches proposed in the literature are typically related to those three challenges. Non-parametric approaches like the analog method (Hamill and Whitaker 2006; Hamill et al. 2015) completely avoid the first two issues, but may be disproportionately affected by the third one since their treatment of high precipitation amounts neglects the information with training samples with lower precipitation amounts. Parametric methods, on the other hand, can extrapolate the relations found between observations and forecasts of low and moderate magnitudes to higher magnitudes. In doing so, they may reduce the demand for training data, but the quality of the corresponding predictions strongly depends on the adequacy of the parametric assumptions that have to be made. Examples of parametric approaches that have been developed for quantitative precipitation forecasts include Bayesian Model Averaging (BMA, Sloughter et al. 2007), extended logistic regression (ExLR, Wilks 2009; Ben Bouallègue 2013; Messner and Mayr 2014), and ensemble model output statistics (EMOS, Scheuerer 2014). All of them make somewhat ad-hoc assumptions about the parametric form of the predictive distributions: Sloughter's BMA method models precipitation occurrence/non-occurrence separately and assumes gamma distributions for positive precipitation amounts; ExLR implies the assumption of censored logistic distributions; Scheuerer's EMOS method assumes censored generalized extreme value distributions. To deal with the issue of heteroscedasticity mentioned above, BMA and ExLR commonly apply power-transformations to both forecasts and observations, with powers chosen such as to make the forecast error terms more homoscedastic. Scheuerer's EMOS method utilizes two different ensemble statistics that serve as as predictors for the scale parameter of the censored GEV distributions.

In this paper we will leverage NOAA's second-generation GEFS reforecast data set (Hamill et al. 2013) to systematically develop a parametric model for the conditional distribution of observed precipitation amounts given the ensemble forecasts. This will eventually lead to an approach similar to the one proposed by Scheuerer (2014), but based on censored, shifted gamma distributions (CSGD), and a more sophisticated heteroscedastic regression model that accounts for some further peculiarities of precipitation. In Section 2 we briefly describe the forecast and observation data used in this study, and we introduce our CSGD model in Section 3. Section 4 describes the actual post-processing approach which proceeds in three steps: first, a CSGD model for the climatological distribution of the observations is fitted; second the ensemble forecasts are adjusted such as to match this observation climatology and are condensed into three ensemble statistics. Finally, a nonhomogeneous regression model is set up which links these statistics to the CSGD parameters, and results in a conditional distribution model for the observations given the ensemble forecasts. This model is relatively complex, but a comparison with non-parametrically estimated conditional distributions of observed precipitation amounts shows that a certain degree of flexibility (and thus complexity) is necessary to address the peculiarities of precipitation. The benefit of developing a sophisticated parametric approach will become clear in Section 5, where probabilistic forecasts generated by our method are verified and compared against those obtained with a state-of-the-art analog approach. The latter is even more flexible and easier to implement, but in situations where training data is sparse (e.g. rare events) the predictive performance of our method is favorable. The issue of limited training sample size is further discussed in Section 6, and plans for future investigations and development are pointed out.

2. Data

The post-processing method developed here are applied to 12-hourly accumulated precipitation forecasts during the period from January 2002 to December 2013 for lead times up to +6 days. All of the forecast data were obtained from the second-generation GEFS reforecast data set; the same data was used in a recent paper by Hamill et al. (2015) which discusses variants of the analog method for statistical post-processing of ensemble precipitation forecasts. For precipitation, individual forecasts by the 11-member GEFS reforecast ensemble were retrieved, and forecast data was extracted on GEFS's native Gaussian grid at $\sim$1/2-degree resolution in an area surrounding the contiguous U.S. Total-column ensemble-mean precipitable water is used as an additional predictor in our regression model, and the corresponding forecasts were interpolated to the same grid before further processing. Again as in Hamill et al. (2015), post-processing and verification is performed against precipitation analyses from the climatology-calibrated precipitation analysis (CCPA) data set of Hou et al. (2014), which were obtained on a $\sim$1/8-degree grid inside the contiguous U.S. The downscaling from the $\sim$1/2-degree to the $\sim$1/8-degree resolution will implicitly be part of the post-processing procedure.

3. The censored, shifted gamma distribution

To set up a parametric post-processing method, a suitable class of probability distributions must be identified. As precipitation occurrence/non-occurrence and amount are modeled jointly, a convenient way to do so is using a continuous distribution that permits negative values, and left-censoring it at zero, i.e. replacing all negative values by zero. The censoring turns the probability for negative values of the uncensored distribution into a probability of observing a value equal to zero, thus ensuring requirement 1 from above.

Exploratory data analysis reveals another challenging requirement for conditional distributions of precipitation accumulations: when the predictor variable (e.g. the ensemble-mean precipitation forecast) is small, then a strongly right-skewed distribution is called for, but the required skewness becomes smaller and smaller as the predictor variable's magnitude increases. To some extent, this behavior can be addressed by using gamma distributions, which are characterized by a shape parameter k and a scale parameter θ. Those two parameters are related to the mean μ and the standard deviation σ of the gamma distribution via

$$k = \frac{\mu^2}{\sigma^2}, \qquad \theta = \frac{\sigma^2}{\mu} \tag{1}$$

(Wilks 2011, Sec. 4.4.3). Since the predictive standard deviation increases more slowly than the predictive mean as the predictor variables increase, the shape parameter k decreases, and with it the skewness of the distribution.

A disadvantage of the gamma distribution is that its value range is non-negative. To make the above censoring idea work, we therefore introduce an additional parameter $\delta > 0$, which shifts the cumulative distribution function (CDF) of the gamma distribution somewhat to the left. That is, if F_k denotes the CDF of a gamma distribution with unit scale and shape parameter k, then the CDF $\tilde{F}_{k,\theta,\delta}$ of our censored, shifted gamma distribution (CSGD) model is defined by

$$\tilde{F}_{k,\theta,\delta}(y) = \begin{cases} F_k\left(\frac{y-\delta}{\theta}\right) & \text{for } y \geq 0 \\ 0 & \text{for } y < 0 \end{cases} \tag{2}$$

Using the relations in (1), this distribution can also be parametrized by μ, σ, and δ: μ reflects the expected magnitude of precipitation; σ parametrizes prediction uncertainty; δ reduces the magnitude of precipitation somewhat and controls the probability of zero precipitation. An illustration of the CSGD is given in Fig. 1.

4. Post-processing method

Having selected a family of probability distributions, we need to set up a model that links the three parameters of this distribution to the ensemble forecasts. This is done in three steps. First, we fit a CSGD model as in eq. (2) to the observed precipitation accumulations at each grid point (separately for each month) to describe the observation climatology. In the second step, quantile mapping is performed to adjust the ensemble precipitation forecasts such as to match this observation climatology. The adjusted forecasts are then reduced to two statistics that measure mean and spread of predicted precipitation accumulations. A further statistic is calculated that measures the mean precipitable water. Finally, a heteroscedastic regression model is set up that links these statistics to the CSGD parameters, and thus yields, for given ensemble forecasts, a predictive distribution for the observed precipitation accumulations.

a. Unconditional precipitation accumulations

Although our main interest is in modeling the conditional distribution of observed precipitation accumulations given the ensemble forecasts, we first consider their unconditional (i.e. climatological) distributions. Studying those is much easier and yet quite instructive, as the conditional distributions should converge towards the unconditional distribution as forecast skill decreases. Moreover, they will allow us to parameterize the conditional distributions such as to make them more comparable across grid points with different climatologies.

To fit the parametric CDF $\tilde{F}_{\mu,\sigma,\delta}$ to the empirical CDF $\hat{F}_n$ of the observations $y_1,\ldots,y_n$ at this grid point, we minimize the integrated quadratic distance

$$d_{IQ}\big(\tilde{F}_{\mu,\sigma,\delta},\hat{F}\big) = \int_0^\infty \big(\tilde{F}_{\mu,\sigma,\delta}(t) - \hat{F}_n(t)\big)^2 dt \tag{3}$$

in μ, σ, and δ. According to Thorarinsdottir et al. (2013), this is equivalent to minimizing the mean continuous ranked probability score (CRPS)

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{crps}\big(\tilde{F}_{\mu,\sigma,\delta}, y_i\big) \tag{4}$$

where

$$\mathrm{crps}(F, y) = \int_{-\infty}^{\infty} \big(F(t) - H(t-y)\big)^2 dt, \tag{5}$$

and $H(\cdot)$ is the Heaviside step function, i.e. it is equal to 1 if $t \geq 0$ and zero otherwise. After re-parameterizing, the integral on the right hand side can be expressed in closed form as

$$\begin{aligned} \mathrm{crps}\big(\tilde{F}_{k,\theta,\delta}, y\big) \;=\; & \theta\tilde{y}\big(2F_k(\tilde{y}) - 1\big) \;-\; \theta\tilde{c}\,F_k(\tilde{c})^2 \\ & + \theta k\big(1 + 2F_k(\tilde{c})F_{k+1}(\tilde{c}) - F_k(\tilde{c})^2 - 2F_{k+1}(\tilde{y})\big) \\ & - \frac{\theta k}{\pi} B\big(\tfrac{1}{2}, k+\tfrac{1}{2}\big)\big(1 - F_{2k}(2\tilde{c})\big) \end{aligned}$$

where $\tilde{c} := -\frac{\delta}{\theta}$, $\tilde{y} := \frac{y-\delta}{\theta}$ and $B(\cdot,\cdot)$ is the beta function (a derivation of this formula is given in the online appendix). The availability of a closed form expression makes model fitting through numerical CRPS minimization computationally efficient. When performing this minimization, the constraint $\delta \geq -\mu$ is imposed in addition to the constraints $\mu, \sigma > 0$ and $\delta \leq 0$ that are required for the distribution model to be well-defined. The reason for this will become more clear later in this section, when we set up the regression model for the conditional distribution of the observation given the forecasts.

For solving the constrained optimization problems numerically, we use the Fortran 77 implementation of the Linearly Constrained Optimization Algorithm (LINCOA) by Michael J. D. Powell (details of this algorithm have not been published yet, but the usual way of choosing a new vector of variables is described in Powell 2014). A starting value for the optimization is obtained through the following rationale: if we had $\sigma = \mu$, the underlying gamma distribution would have

a shape parameter $k = 1$, which corresponds to the special case of an exponential distribution. For this distribution, the mean over all non-zero precipitation amounts is an estimate of μ (and σ), for any probability of precipitation π_{pop}, and δ can subsequently be estimated as $\delta = \mu \log(\pi_{pop})$. For the 12h-accumulations considered here, the best-fitting k is typically smaller than 1, with μ being overestimated by the assumption of an exponential distribution. Moreover, the first guess estimates proposed above might violate the constraint $\delta \geq -\mu$. We therefore improve our first guess by fixing σ, gradually decreasing μ, and recalculating $\delta = -(\mu/k) \cdot F_k^{-1}(1 - \pi_{pop})$ until $\delta > -\mu/2$. The resulting values of μ, σ, and δ are then used as starting values for the numerical minimization of (4). If $\pi_{pop} < 0.02$, we expect the number of days with non-zero precipitation to be too small to warrant stable estimates, and we therefore take the starting values as the final estimates. For extremely dry grid points with $\pi_{pop} < 0.005$, even the simple preliminary estimates might be unreliable, and we use ad-hoc values $\mu = 0.0005, \sigma = 0.0182, \delta = -0.00049$ to set up a parametric distribution model for the analyzed climatology. Figs. 2 and 3 show examples of fitted CSGDs at a very wet location (West Palm Beach, FL) and a very dry location (Phoenix, AZ), respectively. The empirical and the fitted, parametric CDFs are virtually indistinguishable. The approximate character of the parametric distribution becomes more obvious when we plot its quantiles against the sorted observations. In those Q-Q plots we observe quite strong departures from the diagonal, especially in the upper tail. However, this is also where we expect significant sampling variability. In order to understand to what extent the departures might just be random, we add pointwise 95% Monte Carlo intervals by simulating 10000 samples of the same size as the original observations according to the fitted distribution model, sorting them, and reporting the 2.5% and 97.5% quantile of the first elements, second elements, and so forth. The black dots in the Q-Q plots in Fig. 2 and 3 (and in all other examples that we studied) are mostly inside the

95% Monte Carlo intervals, suggesting that the distribution family proposed here is adequate for modeling unconditional distributions of precipitation accumulations.

b. Quantile mapping and ensemble statistics

As a second step in our post-processing scheme, we attempt to correct systematic errors in ensemble forecast climatology. For example, the underlying numerical weather prediction may produce too many days with light precipitation and underforecast heavy precipitation events. Alternatively, these errors can arise due to coarser spatial resolution of the forecast grid compared to the grid on which analyzed precipitation is available. Let s be a location associated with some analysis grid point. Prediction errors of the ensemble forecasts may result from inaccurately predicted magnitudes of a precipitation event as described above, but may also be caused by displacement errors. Following Scheuerer (2014), we therefore consider ensemble forecasts at all forecast grid points within a certain neighborhood $N(s)$ of s as potential predictors for the analyzed precipitation amount at s. Forecast f_{xj} of ensemble member j at forecast grid point x is thus used multiple times to calculate ensemble- and spatial means and spreads for all analysis grid point neighborhoods $N(s_1), N(s_2), \ldots$ containing x. Each time, the climatological adjustment is made with respect to the respective analysis grid point $s_1, s_2, \ldots$ as illustrated in Fig. 4. To match the forecast climatology with the observation climatology, quantile mapping is performed: for each forecast we determine to which quantile of the forecast climatology it corresponds, and then map it to the corresponding quantile of the observation climatology. Formally, denote by $F_{fcst,x}$ the CDF of the forecast climatology at x and by $F^{-1}_{obs,s}$ the quantile function of the observation climatology at s. Then the adjusted forecast $\tilde{f}_{xj}$ of ensemble member j at x is given by

$$\tilde{f}_{xj} := \begin{cases} 0 & \text{if} \quad f_{xj} = 0 \\ F^{-1}_{obs,s}\big(F_{fcst,x}(f_{xj})\big) & \text{if} \quad f_{xj} > 0 \end{cases} \tag{6}$$

The CDF $F_{fcst,x}$ and the quantile function $F^{-1}_{obs,s}$ must be estimated from suitable subsets (to account for seasonal differences) of the training data. If the ensemble members are exchangeable as it is with the GEFS data used here, we can pool the training data over all members and approximate $F_{fcst,x}$ by the empirical CDF of all training forecasts at x. The quantile function $F^{-1}_{obs,s}$ can either be approximated by interpolating the empirical quantiles of the training observations at s. Alternatively, one can invert the CDF of the observation climatology fitted in the first step as described above. For very high quantiles, those two choices can differ noticeably, with the empirical quantiles being subject to substantial sampling variability and the model-based quantiles being subject to possible biases due to the parametric assumption on the form of the distributions (see Figs. 2 and 3). We consider those approximation biases the lesser of two evils and prefer the model-based quantiles, but as a safeguard against unwarranted corrections of high forecast values, $\tilde{f}_{xj}$ is capped at the level $1.3 \cdot F^{-1}_{fcst,x}(0.999)$.

To use these adjusted ensemble forecasts within a regression framework, they need to be condensed into statistics that summarize the most important information. While we think that all forecast grid points in $N(s)$ - which we take as a neighborhood around s with radius $r = 1$ degree - should be considered, we still expect forecasts at grid points closer to s to be more informative about the weather at s. Following Scheuerer (2014), we therefore weigh the forecast grid points according to their distance to s and let

$$w_{sx} \sim \max\left\{1 - \left(\frac{dist(x,s)}{r}\right)^2, 0\right\}$$

with a constant of proportionality chosen such that the weights sum up to one (see Fig. 4 for an illustration of this weighting scheme). Assuming that we have adjusted precipitation forecasts $\tilde{f}_{x1}, \ldots, \tilde{f}_{xm}$ and forecasts $\chi_{x1}, \ldots, \chi_{xm}$ of precipitable water, we consider the following ensemble

statistics:

$$\overline{f}_s := \frac{1}{m}\sum_{j=1}^{m}\sum_{x\in N(s)} w_{sx}\,\tilde{f}_{xj} \tag{7}$$

$$\overline{\chi}_s := \frac{1}{m}\sum_{k=1}^{m}\sum_{x\in N(s)} w_{sx}\,\chi_{xk} \tag{8}$$

$$\mathrm{MD}_{f,s} := \frac{1}{m^2}\sum_{j,j'=1}^{m}\sum_{x,x'\in N(s)} w_{sx}w_{sx'}\,|\tilde{f}_{xj}-\tilde{f}_{x'j'}| \tag{9}$$

The first two are the weighted means of predicted adjusted precipitation accumulations and precipitable water over all ensemble members and all forecast grid points in $N(s)$. The third statistic measures the dispersion of the predicted precipitation accumulations both between ensemble members and between grid points in $N(s)$. Unlike Scheuerer (2014), we do not use separate measures of dispersion for those two sources of variability in order to keep the number of parameters in our heteroscedastic regression model (defined below) as small as possible. We finally note that the adjustment of forecasts in the neighborhood of s to the analysis climatology at s via quantile mapping achieves two goals: first an implicit downscaling to a finer grid, and second the retention (or even enhancement) of orographically related features in the raw ensemble forecasts when averaging over $N(s)$. The latter is illustrated in Fig. 3 of Scheuerer (2014) where a simpler, multiplicative adjustment is used for that purpose.

c. Regression equations

The final step is now to set up and fit a regression model for the conditional distribution of observed precipitation accumulations given the forecasts. To this end, the ensemble statistics for location s defined above must be linked to the parameters μ_s, σ_s, and δ_s of our CSGD model in eqs. (1) and (2). Denote by $\mu_{cl,s}, \sigma_{cl,s}$ and $\delta_{cl,s}$ the parameters of the climatological CSGD at s. We model the conditional CSGDs as deviations from the climatological CSGD via the following

equations

$$\mu_s = \frac{\mu_{cl,s}}{\alpha_{1,s}} \cdot \log\left(1 + \alpha_{1,s}\left[\alpha_{2,s} \cdot \left(\exp(1) - 1\right) + \alpha_{3,s} \cdot \frac{\overline{f}_s}{\overline{f}_{cl,s}} + \alpha_{4,s} \cdot \frac{\overline{\chi}_s}{\overline{\chi}_{cl,s}}\right]\right) \tag{10}$$

$$\sigma_s = \alpha_{5,s} \cdot \sigma_{cl,s} \cdot \sqrt{\frac{\mu_s}{\mu_{cl,s}}} + \alpha_{6,s} \cdot \mathrm{MD}_{f,s} \tag{11}$$

$$\delta_s = \delta_{cl,s} \cdot \left(\alpha_{7,s} + \alpha_{8,s} \cdot \frac{\mu_s}{\mu_{cl,s}}\right) \tag{12}$$

where $\overline{f}_{cl,s}$ and $\overline{\chi}_{cl,s}$ denote the climatological means of $\overline{f}_s$ and $\overline{\chi}_s$, respectively, calculated as averages of these quantities over the current training sample.

The form of the regression equations (10)-(12), which depend on the fitted parameters $\alpha_{1,s}, \ldots, \alpha_{8,s}$, require some explanation. Consider first a situation with very good predictability. In this case, we often have $0 < \alpha_{1,s} \ll 1$, which implies $\log(1 + \alpha_{1,s} z) \approx \alpha_{1,s} z$, and reduces eq. (10) to a linear regression on the two multiplicatively normalized predictors $\overline{f}_s$ and $\overline{\chi}_s$. Eq. (11) accounts for the heteroscedasticity in the uncertainty about precipitation accumulations in two different ways. The first term increases σ_s proportionally to the square root of μ_s, which accounts for the fact that forecast uncertainty increases with the magnitude of expected precipitation amounts. The second term is proportional to $\mathrm{MD}_{f,s}$ and thus accounts for flow-dependent uncertainty. Eq. (12) permits an increased shift with increasing μ_s. This is useful to address one shortcoming of the CSGD when it is used as a model for conditional distributions of the observed given the forecasts: the CSGD yields very good fits for low to moderate levels of predicted precipitation, but for elevated levels its left tail can become too light. In that situation, increasing δ_s proportionally to μ_s permits a certain degree of re-adjustment of the lower predictive quantiles as can be seen by comparing the probability density functions in Fig. 5. Another peculiarity of conditional distributions for precipitation is that a linear increase of μ_s with $\overline{f}_s$ does not always seem adequate. Especially in situations with reduced predictability (longer lead times, summer season), ensemble forecasts of high precipitation amounts are often unreliable and should be decreased proportionately more

compared to forecasts of intermediate levels. This is the rationale behind the logarithm in eq. (10). Increasing the parameter $\alpha_{1,s}$ reduces the growth of μ_s with increasing predictors and thus accounts for the phenomenon just described. Fig. 5 illustrates the evolution of the predictive CSGD density with increasing mean precipitation $\overline{f}_s$ in a simplified setting where $\alpha_{4,s}$ and $\alpha_{6,s}$ have been set to zero. It shows how both uncertainty and shift parameter increase with increasing $\overline{f}_s$; at the same time the skewness of the underlying gamma distribution becomes smaller and smaller. Choosing $\alpha_{1,s} = 0.05$ results in a moderate departure from a linear relation between $\overline{f}_s$ and μ_s.

Is the CSGD adequate for modeling conditional distributions of precipitation accumulations, and are the above regression equations for its parameters μ, σ, and δ adequate for describing the evolution of these parameters with increasing ensemble mean? To answer this we compare quantiles derived from predictive CSGDs with empirical conditional quantiles obtained without any parametric assumption. For this purpose, however, even the 12 years worth of reforecast data are not enough if only data from a single grid point are considered. We focus on the analysis grid point corresponding to the city of Atlanta, GA, and we increase the corresponding dataset by selecting 200 additional grid points within a radius of about 700 km around Atlanta that have a similar climatology and are at least 40 km apart from each other. For each season, we then have about $91 \times 12 \times 201$ pairs of observations and quantile adjusted forecasts. We study again the simplified regression model with $\alpha_{4,s} = \alpha_{6,s} = 0$, i.e. with $\overline{f}_s$ as the only predictor. The conditional quantiles of the observation given $\overline{f}_s = x$ can then be approximated by considering all forecast-observation pairs for which $\overline{f}_s$ falls within a certain window $(x-\varepsilon, x+\varepsilon)$ around the precipitation amount x, and computing the quantiles of the corresponding observations. We let ε increase with x to account for the fact that the number of pairs with $\overline{f}_s \approx x$ decreases rapidly as x increases. For $x =$5 mm and $x =$15 mm our choice of ε is illustrated in Fig. 6. The crosses in each plot correspond to the empirical, conditional deciles (i.e. quantiles for the probabilities $0.1, \cdots, 0.9$)

for each season and forecast lead times +12 to +24 h and +108 to +120 h. The solid lines are the quantiles obtained with our parametric regression model, fitted to the same training data. This is done again by CRPS minimization using the LINCOA algorithm by Michael J. D. Powell. Clearly, not every model-based quantile approximates the respective empirical quantile perfectly, and very irregular behavior cannot be captured. Yet one can see that the increase of predictive uncertainty with increasing $\overline{f}_s$ is captured quite well; further the non-linear relation between $\overline{f}_s$ and μ_s, which takes different forms depending on the skill of the ensemble forecasts, is accounted for by our regression model. It is worth noting that our method for getting empirical estimates of conditional quantiles is quite similar to what is being done by analog approaches. Those techniques are much more flexible and avoid the approximation errors entailed by parametric methods. On the other hand, several of the plots in Fig. 6 also suggest that the empirical quantiles for large values of $\overline{f}_s$ are subject to quite substantial sampling error, even in the situation considered here where we choose the "analogs" from a training data set of size $91\times12\times201$.

Finally, consider how the regression model (eqs. (10)-(12)) for the predictive CSGDs approaches the parameters for the climatological CSGD in the limit where the raw ensemble forecasts have no skill. As the lead time increases, one can expect that the three predictors $\overline{f}_s, \overline{\chi}_s$ and $\mathrm{MD}_{f,s}$ become less and less informative about the true weather, and so the corresponding regression parameters $\alpha_{3,s}$, $\alpha_{4,s}$, and $\alpha_{6,s}$ tend to zero. We already mentioned that $\alpha_{1,s}$ typically increases with decreasing skill of the ensemble, and we choose $\alpha_{1,s} \leq 1$ as an ad-hoc upper bound. The decrease of $\alpha_{3,s}$ and $\alpha_{4,s}$ goes along with an increase of the intercept parameter $\alpha_{2,s}$, and for $\alpha_{2,s} = 1$, we end up with $\mu_s = \mu_{cl,s}$. Decreasing forecast skill also entails increasing prediction uncertainty, and we retrieve the climatological value $\sigma_s = \sigma_{cl,s}$ as $\alpha_{5,s}$ tends to 1. For $\alpha_{7,s}$ and $\alpha_{8,s}$, there is no obvious tendency, and as μ_s approaches $\mu_{cl,s}$, they also become less and less identifiable. As long as their sum tends to 1, however, δ_s approaches $\delta_{cl,s}$, and the climatological CSGD results as a limiting

case. Including $\mu_{cl,s}, \sigma_{cl,s}$ and $\delta_{cl,s}$ in our parametrization therefore helps reducing the dependence of the regression parameters $\alpha_{1,s}, \ldots, \alpha_{8,s}$ on the climatology at location s and thus renders them more comparable across different gridpoints.

Modeling the conditional distributions as deviations from the climatological distributions requires some constraints of the latter. We found that this deviation concept does not work well at very dry locations if the shift parameter $\delta_{cl,s}$ of the climatological CSGD is large compared to $\mu_{cl,s}$. In this case, positive precipitation accumulations correspond to the tail end of the underlying gamma distribution, and deforming this distribution into a CSGD with a moderate to high probability of precipitation is rather unnatural. By introducing the constraint $\delta_{cl,s} \geq -\mu_{cl,s}$ on the climatology parameters in subsection a), we enforce a very small shape parameter k. The mass of the underlying gamma distribution is then concentrated near zero, and a very small shift is sufficient to obtain a high probability of values less than zero. Fitting a climatological CDF to the analysis data under this constraint can result in a slightly sub-optimal fit to the empirical, climatological CDF near zero, but this degradation is offset by the fact that the fitted CSGD permits a natural deformation into the predictive CSGD for any value of the predictors.

5. Comparison against the analog method

We apply our CSGD regression method to the full data set described in Section 2. Now, every grid point of the CCPA grid (within the CONUS) is processed separately. Forecasts are cross-validated; for example, 2002 forecasts are trained using 2003-2013 data. In order to account for seasonal differences, a separate set of (both climatological and regression) parameters is fitted for each month; training data is composed of all forecasts and observations from ± 45 days around the 15th of the month under consideration and all years except the one for which forecasts are sought. This results in a training sample size of 91×11 at each grid point. Compared to the

amount of training data that is typically used for weather variables like wind speed or temperature, this training sample size seems fairly large. At very dry locations, however, the majority of both forecasts and observations are zero, and thus carry only limited information that can be leveraged for model fitting. For the observation climatology parameters, the method for proceeding in these dry cases has already been described in Section 4. For the regression parameters, we increase the training data set of any grid point where the climatological probability of precipitation is less than 0.05 by considering also the data at adjacent grid points in east-west and north-south direction. For grid points with a climatological probability of precipitation of less than 0.02, we additionally add the training data from diagonal neighbors. Parameters are estimated via CRPS minimization, subject to the following bounds:

$$\begin{aligned} 0.01 \leq \alpha_{1,s} \leq 1 \quad &, \quad 0 \leq \alpha_{2,s} \leq 1, \\ 0 \leq \alpha_{3,s} \leq 2 \quad &, \quad 0 \leq \alpha_{4,s} \leq 2, \\ 1.2 \leq \alpha_{5,s} \leq 1 \quad &, \quad 0 \leq \alpha_{6,s} \leq 1.5, \\ 0 \leq \alpha_{7,s} \leq 1 \quad &, \quad 0 \leq \alpha_{8,s} \leq 1, \end{aligned}$$

which are partly ad hoc and partly based on the discussion at the end of the previous section.

The predictive performance of the approach presented here is compared to a variant of the "rank analog" approach by Hamill and Whitaker (2006), fully described in Section 2b of Hamill et al. (2015).

a. Brier skill scores

As a measure of predictive performance relative to climatological forecasts, we first consider Brier skill scores computed in the conventional way (Wilks 2011, eqs. 7.34 and 7.35). Scores

for the three thresholds 1, 10, and 25 mm 12 h^{-1} and different forecast lead times are shown in Fig. 7. Even for the >1 mm 12 h^{-1} event, the CSGD method proposed here scores slightly better than the rank-analog method. This is a rather common event at most grid points, and we should expect that sufficiently close analogs can usually be found. The fact that our method can compete with the analog approach in this situation suggests that our parametric approximation does not degrade predictive performance even when analog methods can be expected to perform very well. Comparing results for higher thresholds, we find that the probabilistic CSGD forecasts noticeably improve upon the forecasts by the rank analog method. The event >25 mm 12 h^{-1} is relatively rare even at rather wet grid points, making it difficult to find a sufficient number of suitable analogs. A parametric method, on the contrary, can extrapolate relations found for more common situations and thus yield superior predictions of rare events.

b. Reliability diagrams

To provide some understanding about the causes of the better performance of our parametric method compared to the non-parametric analog approach, we study reliability diagrams for the same events as above (thresholds 1, 10, and 25 mm 12 h^{-1}) and lead times +12 to +24 h and +108 to +120 h. The plots in Figs. 8 and 9 suggest that neither of the two methods are perfectly reliable, but both methods yield probability forecasts that sufficiently accurate. By comparing the inset frequency histograms, one can see that the performance gain of our CSGD method is mainly due to increased resolution. High probabilities for observing heavy precipitation are issued much more frequently, but this is done without degrading the reliability compared to the more flexible analog approach.

c. Case study

We illustrate the last point by considering a heavy precipitation event that took place over the north-western CONUS between 1200 UTC on November 6 and 0000 UTC on November 7 in 2006. Fig. 10 shows the analyzed precipitation accumulations for that period, as well as $+12$ to $+24$ h lead predicted probabilities for exceeding 25 mm 12h^{-1} of precipitation by the raw ensemble, the rank-analog approach and the CSGD regression method. The raw ensemble forecasts for that day were extremely good, but since this is not always the case, one can expect that calibrated probabilistic forecasts modulate the high forecast probabilities. The analog approach modulates them quite strongly, issuing rather moderate probabilities. On the other hand, the CSGD method largely retains the strong signal from the raw ensemble, and hence provides decision makers with a more unequivocal expectation of heavy precipitation.

6. Discussion

We have presented a parametric ("CSGD") post-processing approach that turns statistics of the raw ensemble forecasts into full predictive distributions. Exploratory analysis (see Fig. 2, 3, and 6) showed that censored, shifted, gamma distributions can approximate both climatological distributions of observed precipitation and distributions conditional on the ensemble forecasts reasonably well. Statistics of the ensemble forecasts were calculated summarizing the mean and spread between different ensemble members and different grid points around the location of interest; ensemble mean precipitable water is used as a further predictor. These statistics drive a heteroscedastic regression model, which was demonstrated to be capable of modeling the relation between ensemble forecasts and parameters of the predictive CSGDs. Verification results showed that the CSGD regression approach yields probabilistic forecast that are sufficiently reliable at all lead times, and have better resolution than the forecasts obtained by a state-of-the-art analog ap-

proach. This is especially true for forecasts of extreme events, which are of particular interest due to their socio-economic impact.

The benefits of considering forecasts within a larger neighborhood of the location of interest as predictors have already been demonstrated by Scheuerer (2014), but both in his and in the present paper, the size of this neighborhood was chosen somewhat ad hoc. Ideally, the neighborhood would depend on the particular weather situation; more realistically one could at least vary its size according to the season and lead time. In the slightly different context of choosing a search region for the pattern matching performed as part of the analog method, Hamill and Whitaker (2006) found that the optimal search region increases with lead time. We plan experiments along the same lines in the framework of our CSGD regression approach.

In the present study, the training sample size was about 1000 at each grid point. This appears to be a lot, but for a weather variable like precipitation, where regression relations are somewhat involved and the values of most of the training forecasts and observation are small or moderate, it is not clear if one can fit a sufficiently flexible parametric approach with much less training data. In practice, however, a large reforecast data set like the one used in this study is not always available, and the question arises just how much data is required to warrant stable model fitting and results in reliable forecasts. We will address this question in a separate paper where we will study the effects of training sample size on predictive performance, propose a strategy to overcome possible challenges, and make our CSGD regression method work with a more modest reforecast data set. The results presented here make us confident that a well-designed parametric method is able to provide reliable and sharp probabilistic forecast guidance based on an affordable amount of reforecasts.

References

Ben Bouallègue, Z., 2013: Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Wea. Forecasting*, **28**, 515–524.

Charron, M., G. Pellerin, L. Spacek, P. L. Houtekamer, N. Gagnon, H. L. Mitchell, and L. Michelin, 2010: Toward random sampling of model error in the canadian ensemble prediction system. *Mon. Wea. Rev.*, **138**, 1877–1901.

Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. G. Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast data set. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565.

Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632.

Hamill, T. M., M. Scheuerer, and G. Bates, 2015: Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon. Wea. Rev.*, submitted.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229.

Hou, D., and Coauthors, 2014: Climatology-calibrated precipitation analysis at fine scales: statistical adjustment of stage IV toward CPC gauge-based analysis. *J. Hydrometeor.*, **15**, 2542–2557.

Houtekamer, P. L., and J. Derome, 1995: Methods for ensemble prediction. *Mon. Wea. Rev.*, **123**, 2181–2196.

Messner, J. W., and G. J. Mayr, 2014: Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Mon. Wea. Rev.*, **142**, 448–456.

Molteni, F., R. Buizza, T. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **125**, 73–119.

Powell, M. J. D., 2014: On fast trust region methods for quadratic models with linear constraints. Tech. Rep. DAMTP 2014/NA02, Department of Applied Mathematics and Theoretical Physics, Cambridge University.

Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.*, **140 (680)**, 1086–1096.

Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209–3220.

Thorarinsdottir, T. L., T. Gneiting, and N. Gissibl, 2013: Using proper divergence functions to evaluate climate models. *SIAM/ASA J. Uncert. Quant.*, **1 (1)**, 522–534.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: the generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297–3319.

Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.*, **16**, 361–368.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed., Elsevier Academic Press.

Wilks, D. S., and T. M. Hamill, 2007: Comparison of ensemble-mos methods using gfs reforecasts. *Mon. Wea. Rev.*, **135**, 2379–2390.

LIST OF FIGURES

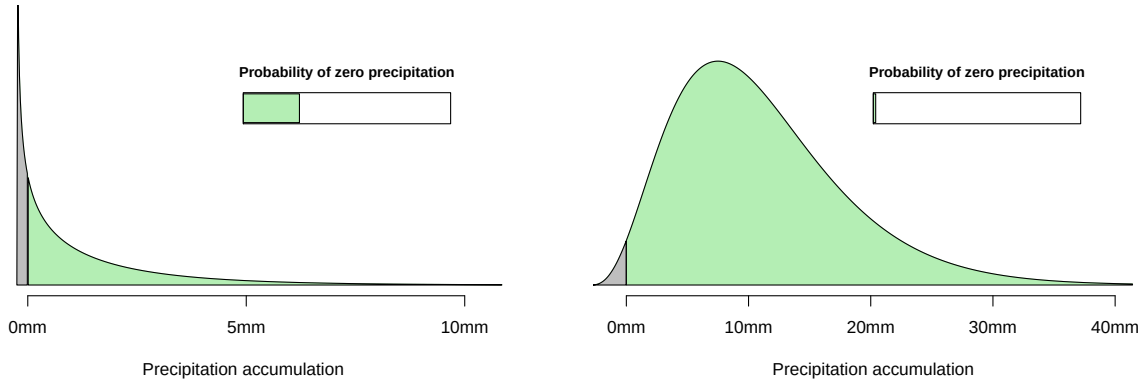


FIG. 1. Examples of censored, shifted gamma distributions. The fractions of the probability density function that fall below zero (shown in the grey shading) translate into a positive probability of being exactly zero.

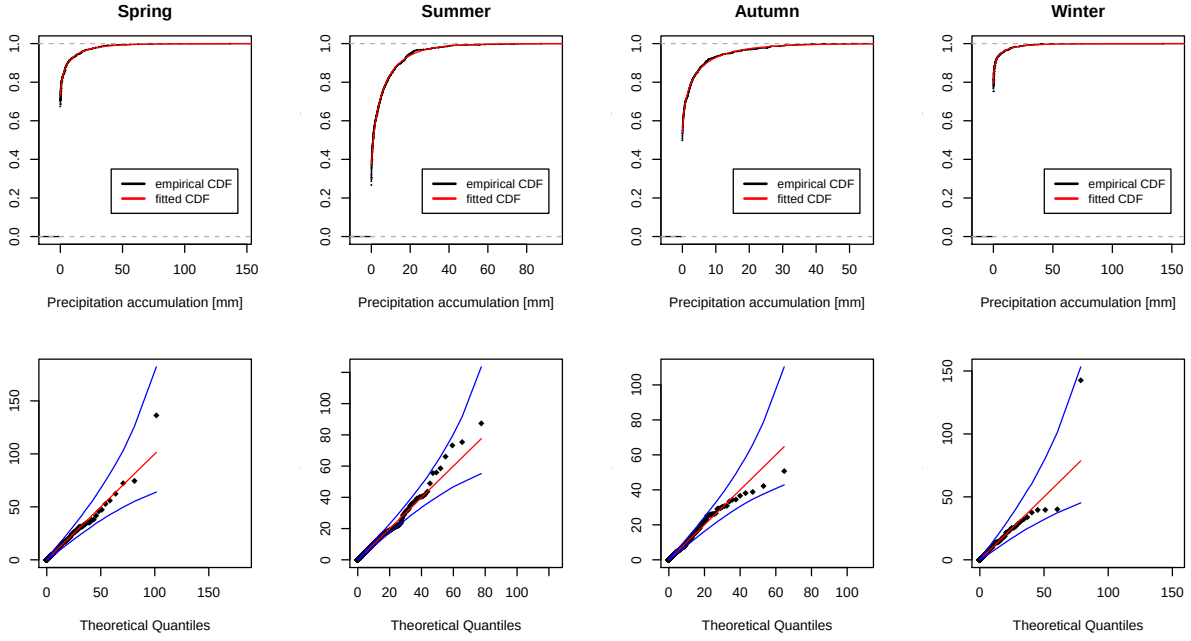


FIG. 2. Empirical and fitted CDFs (top) and Q-Q plots (bottom) of +12 to +24 h analyzed precipitation accumulations in West Palm Beach, FL. The black dots in the lower panels are the sorted observations, plotted against the corresponding theoretical quantiles from the fitted CSGD model. Ideally, they would lie on the diagonal (solid red line); due to sampling variability, however, any black dot lying within the pointwise 95% Monte Carlo intervals (solid blue lines) can still be considered consistent with the fitted model.

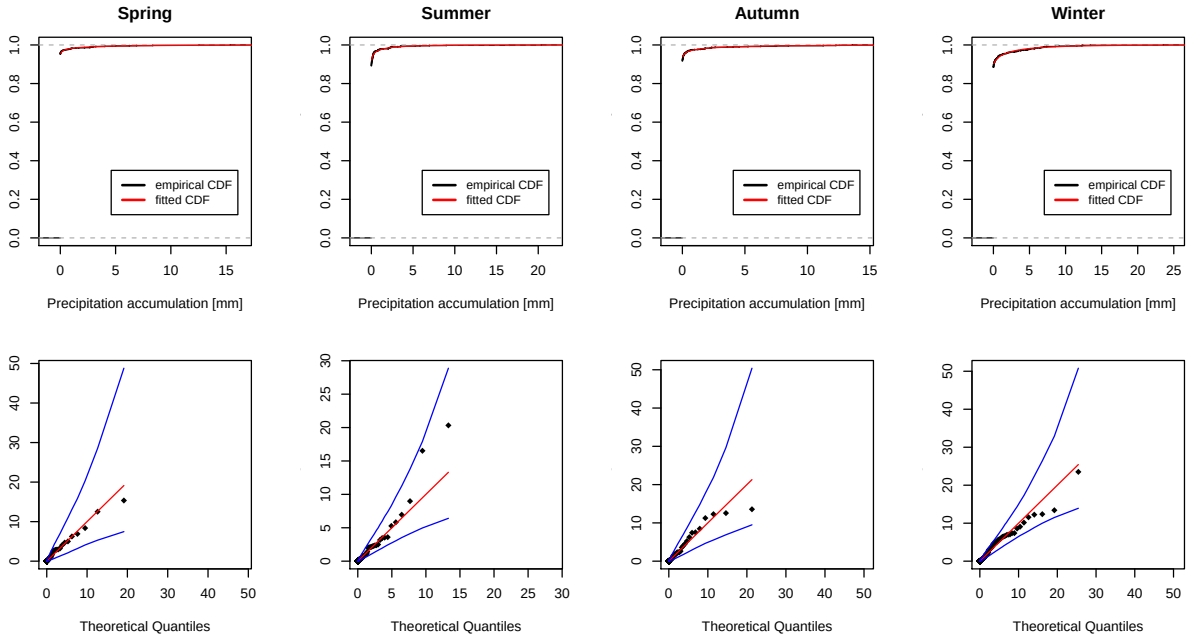


FIG. 3. Same as Fig. 2, but for Phoenix, AZ.

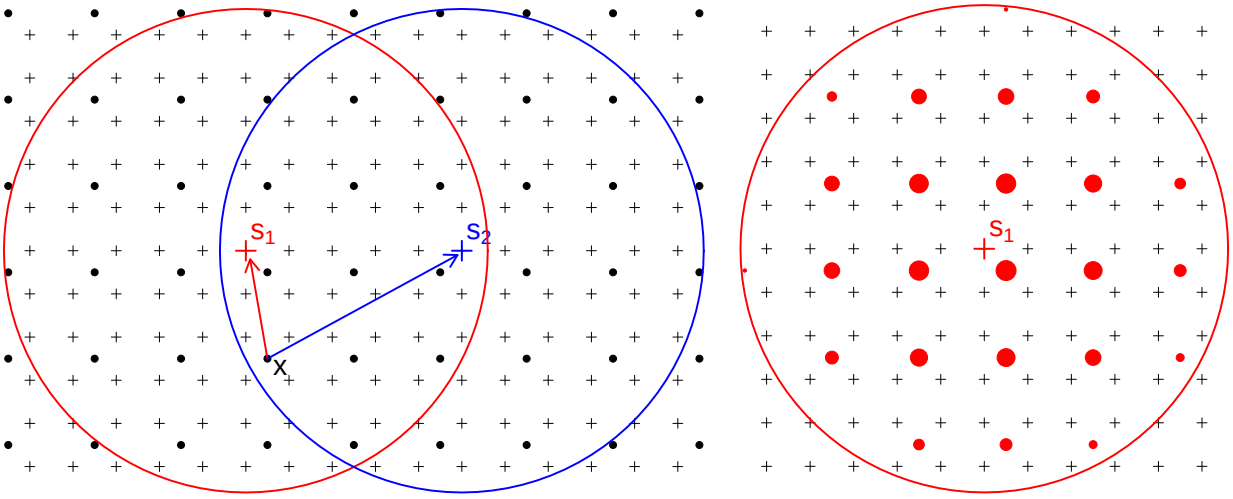
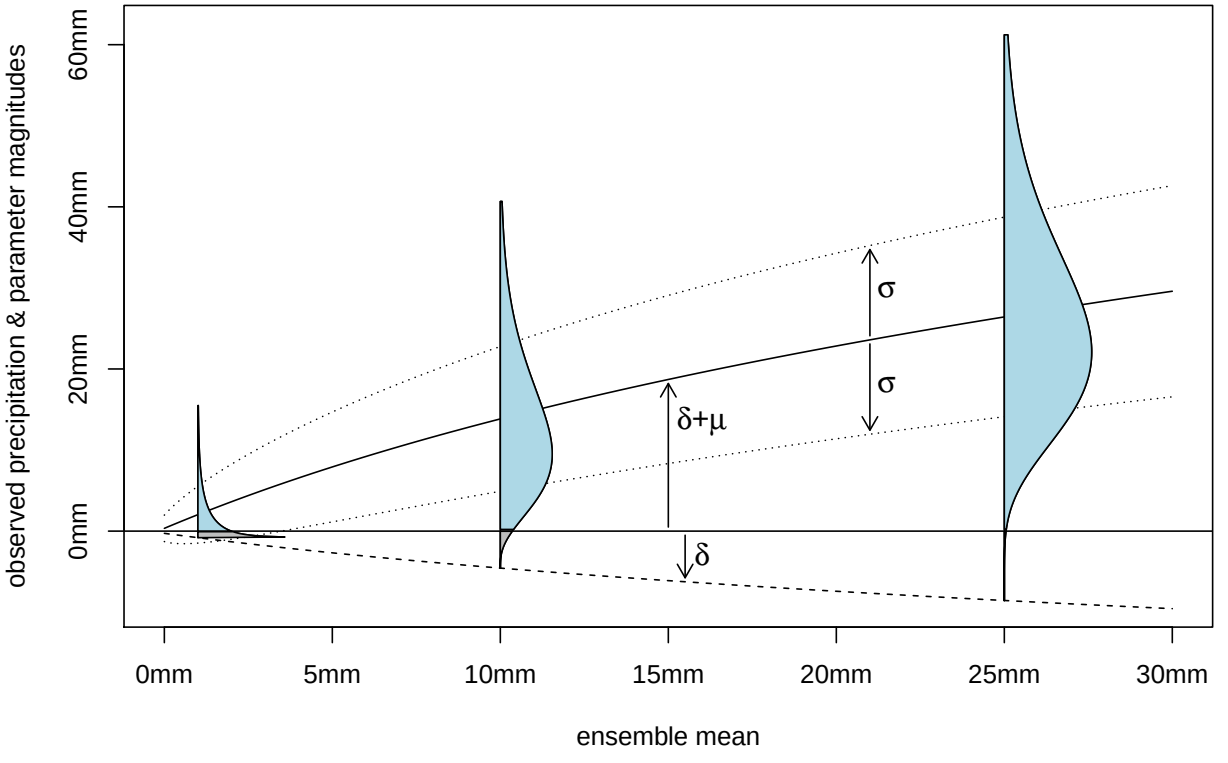


FIG. 4. Illustration of the climatology adjustment (left) and the neighborhood weighting scheme (right). Forecast grid points are denoted by '•', analysis grid points are denoted by '+', and for two of these analysis grid points s_1, s_2 the neighborhoods are delineated by circles. Forecasts at x are used as predictors for s_1, s_2, and other nearby analysis grid points; they are adjusted according to eq. (6) to the respective observation climatologies. The weight applied to a particular forecast grid point within the neighborhood of s_1 is illustrated by the area of the red dots in the right-hand panel.



523 FIG. 5. Example of predictive CSGD densities, showing the evolution of the CSGD parameters μ, σ, and δ
524 from eqs. (1) and (2) as a function of the ensemble-mean statistic $\overline{f}_s$.

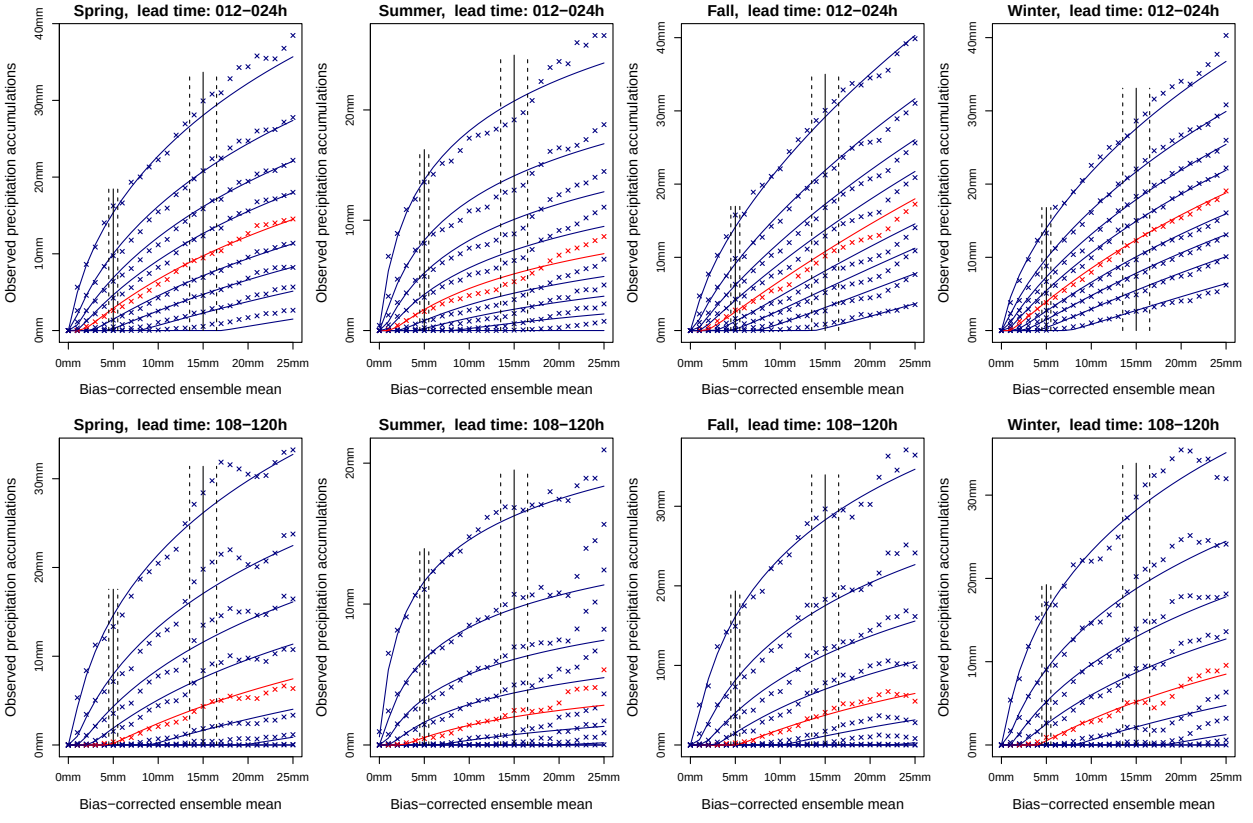
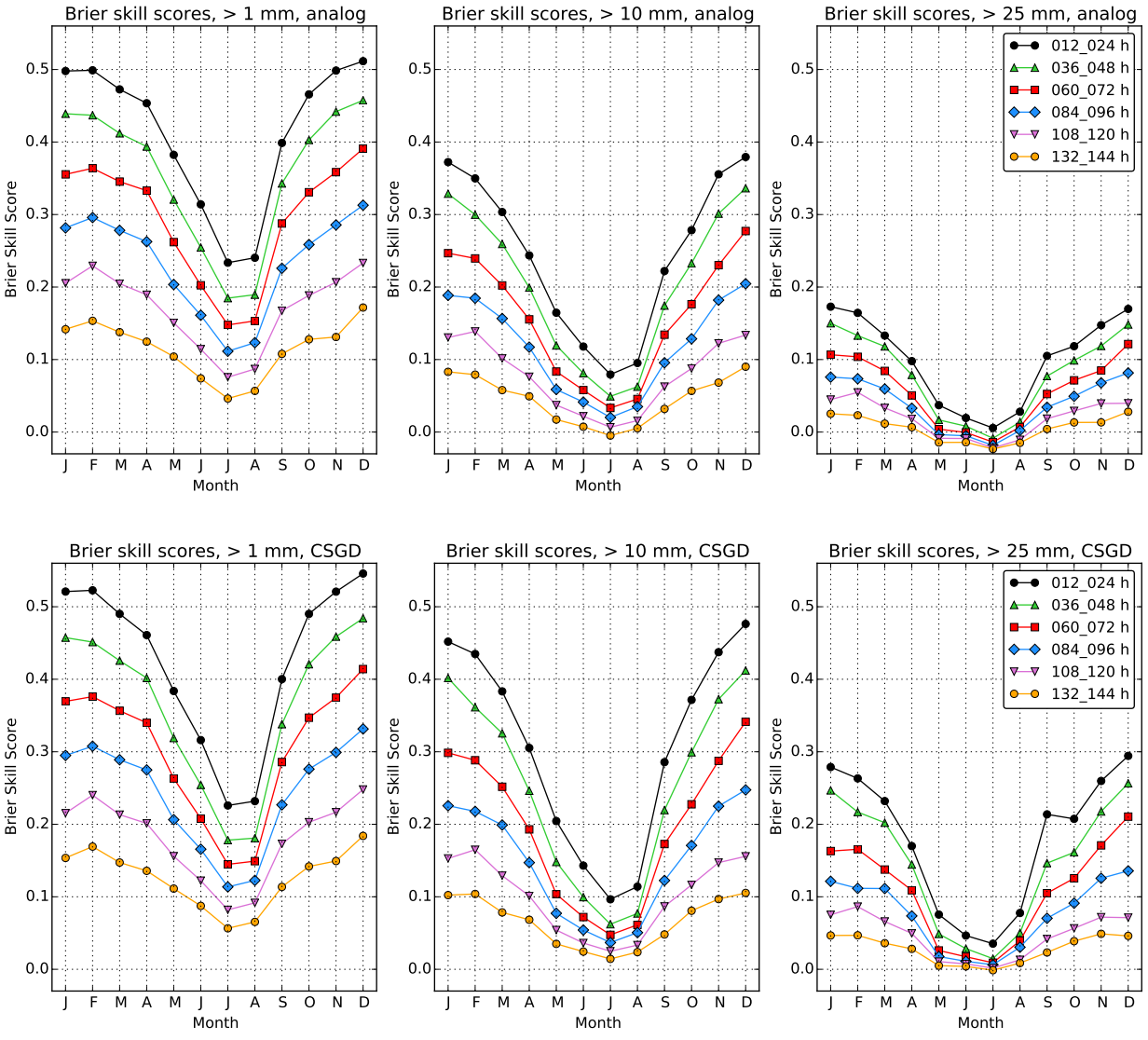
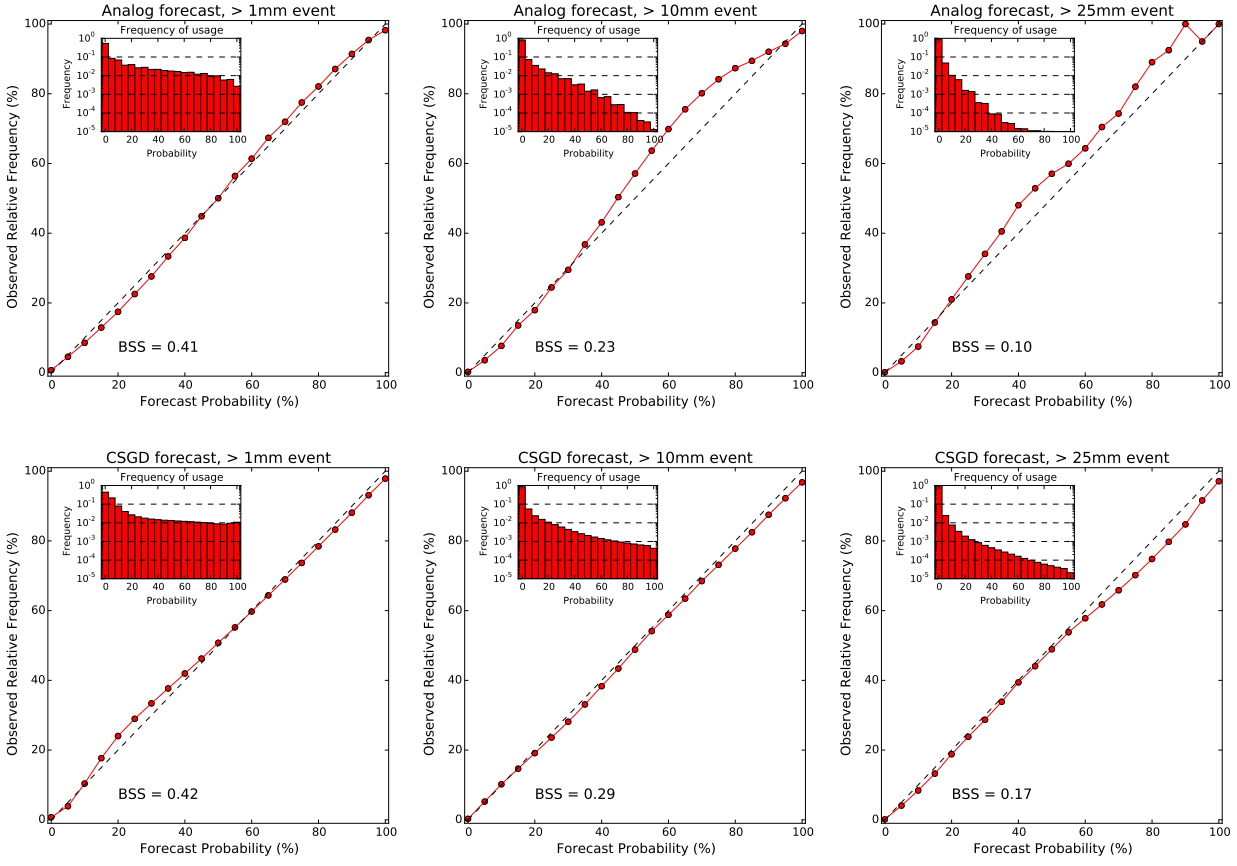


FIG. 6. Conditional deciles (median is highlighted in red) obtained with the augmented Atlanta data set. Empirical deciles are depicted as crosses. For each conditioning value 1 mm, 2 mm, ..., 25 mm they are obtained as empirical deciles of the observations corresponding to ensemble-mean statistics within a certain bin (for 5 mm and 15 mm depicted as vertical dashed lines) around ths value. Deciles derived from the CSGD regression model are depicted as solid lines.



530 FIG. 7. Brier skill scores for different lead times and different event thresholds, separately for each month.
531 Results for the rank-analog method are shown in the top row, those for the CSGD regression approach are shown
532 in the bottom row.



533 FIG. 8. Reliability diagrams for +12 to +24 h lead time and different event thresholds. The top row shows
534 results for the rank-analog method, the bottom row shows results for the CSGD regression approach. The inset
535 histograms depict the frequency with which each category was predicted.

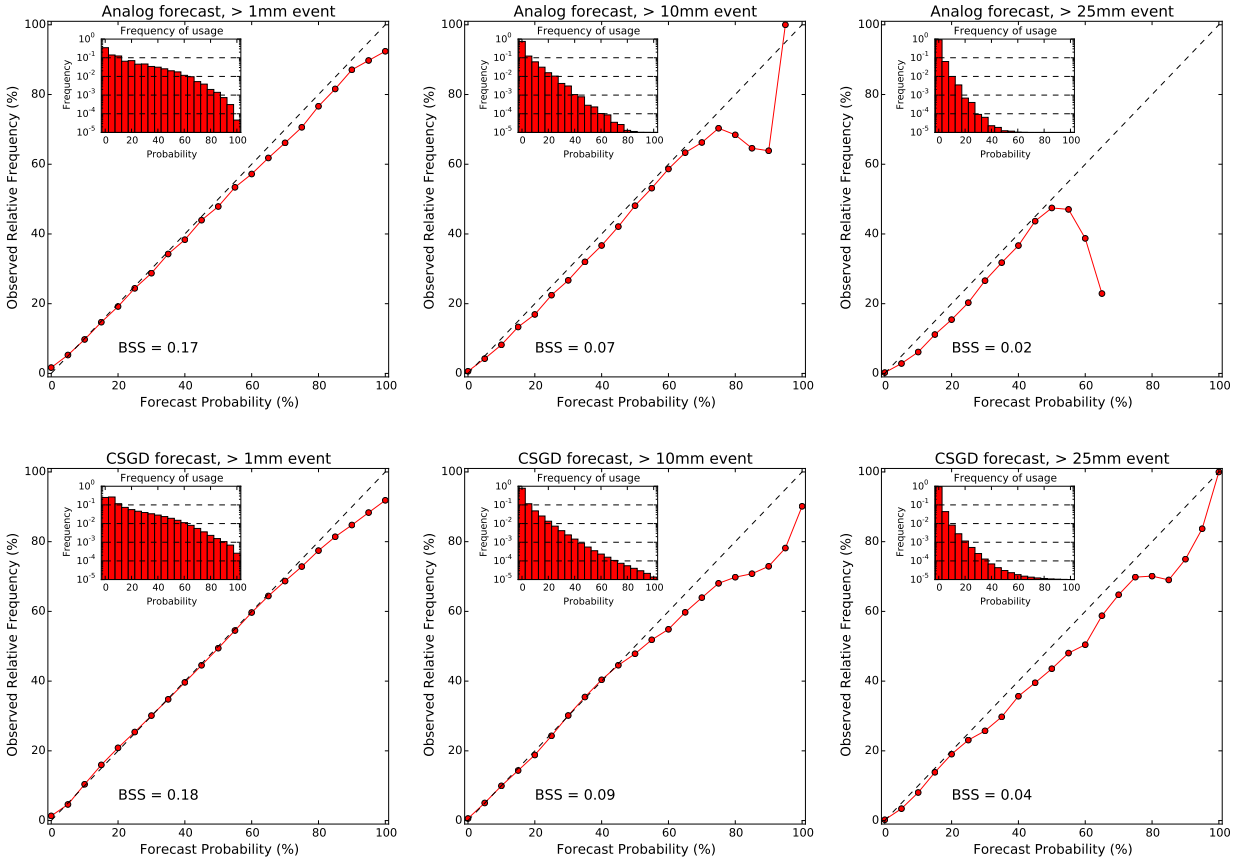


FIG. 9. As Fig. 8 but for +108 to +120 h lead time.

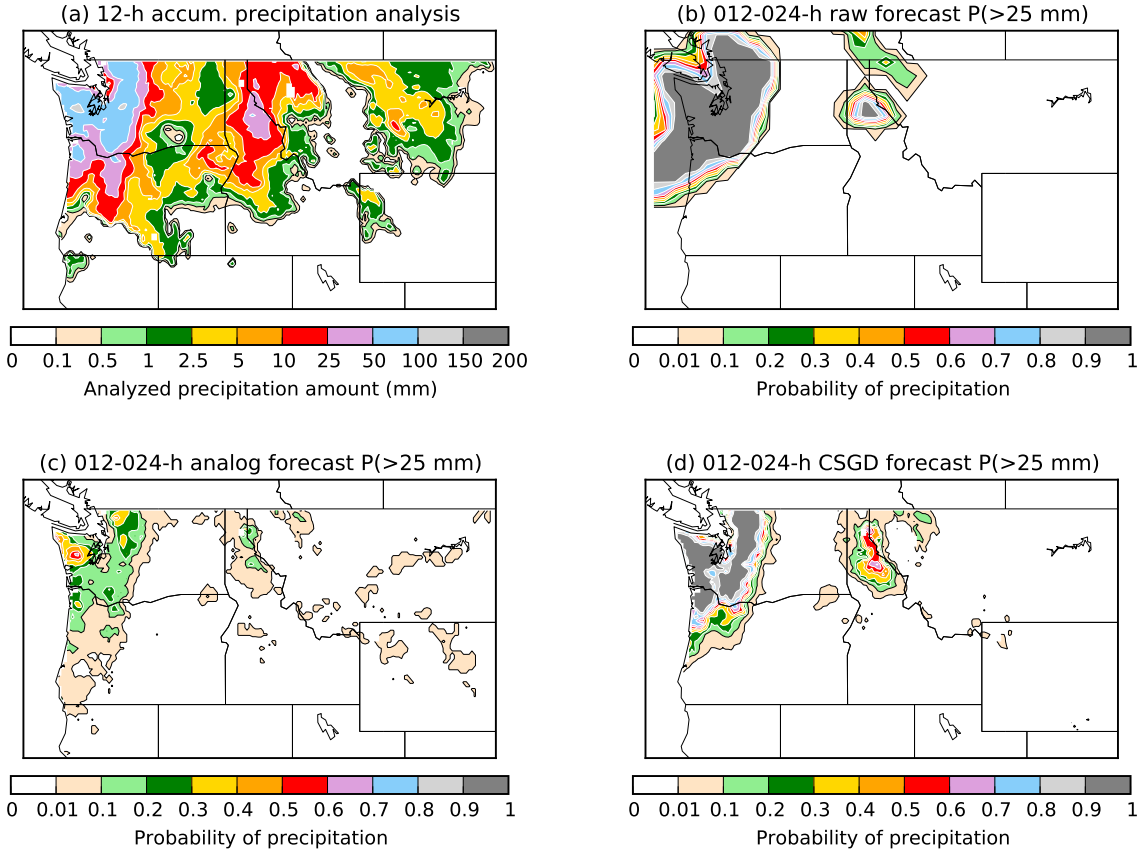


FIG. 10. Analyzed precipitation between 1200 UTC, Nov 6, 2006 and 0000 UTC Nov 7, 2006 (a) and corresponding +12 to +24 h lead probability forecasts for exceeding 25 mm 12 h^{-1} of precipitation by the raw ensemble (b), the rank-analog method (c) and the CSGD regression approach (d).